Knowledge Editing for Large Language Models: A Survey

Year: 2023 | Citations: 196 | Authors: Song Wang, Yaochen Zhu, Haochen Liu

---

Abstract

Large Language Models (LLMs) have recently transformed both the academic and industrial landscapes due to their remarkable capacity to understand, analyze, and generate texts based on their vast knowledge and reasoning ability. Nevertheless, one major drawback of LLMs is their substantial computational cost for pre-training due to their unprecedented amounts of parameters. The disadvantage is exacerbated when new knowledge frequently needs to be introduced into the pre-trained model. Therefore, it is imperative to develop effective and efficient techniques to update pre-trained LLMs. Traditional methods encode new knowledge in pre-trained LLMs through direct fine-tuning. However, naively re-training LLMs can be computationally intensive and risks degenerating valuable pre-trained knowledge irrelevant to the update in the model. Recently, Knowledge-based Model Editing (KME), also known as Knowledge Editing or Model Editing, has attracted increasing attention, which aims at precisely modifying the LLMs to incorporate specific knowledge, without negatively influencing other irrelevant knowledge. In this survey, we aim at providing a comprehensive and in-depth overview of recent advances in the field of KME. We first introduce a general formulation of KME to encompass different KME strategies. Afterward, we provide an innovative taxonomy of KME techniques based on how the new knowledge is introduced into pre-trained LLMs, and investigate existing KME strategies while analyzing key insights, advantages, and limitations of methods from each category. Moreover, representative metrics, datasets, and applications of KME are introduced accordingly. Finally, we provide an in-depth analysis regarding the practicality and remaining challenges of KME and suggest promising research directions for further advancement in this field.